Särtryck ur: Årsbok 2015 KVHAA
Stockholm 2015 (ISBN 978-91-7402-441-8, ISSN 0083-6796)

MATS MALM

Digitala texter och forskningsfrågor

De digitala resurserna har blivit integrerade i vår vardag på ett sätt som knappast hade kunnat förutses för bara tio år sedan. Men samtidigt har vi bara börjat utnyttja potentialerna av dessa resurser för forskningen: vi har en spännande utveckling framför oss. Detta är kanske särskilt tydligt när det kommer till textdatabaser. Å ena sidan har vi fått riklig tillgång till texter – i synnerhet inom de större språkområdena. Å andra sidan finns nu möjligheterna att utveckla helt nya tekniker och metoder som gör litteraturen och inte minst skönlitteraturen till ett inte bara fruktbart utan också hanterligt källmaterial för en rad humanistiska och samhällsvetenskapliga discipliner. Jag skall börja med att presentera digitaliseringsinitiativet *Litteraturbanken* för att sedan resonera kring ett antal lovande vägar att utnyttja materialet för forskningsändamål.

Litteraturbanken uppstod som idé i slutet av 1980-talet vid oasen Timimoun i Sahara. Där satt då författaren Sven Lindqvist och drömde om ett bärbart klassikerbibliotek som kunde rymmas på disketter och i hans ordbehandlare:

> Hit skulle jag alltid vilja komma tillbaka. Här skulle jag tillbringa vintrarna på hotell El Gourara med min ordbehandlare och ett litet diskettbibliotek

med den moderna egoismens klassiker från Hobbes till Huysmans. Och snart nog alla andra texter on line från Europas samtliga nationalbibliotek och databaser. Och så någon gammal bortkastad, häftad och tummad ökenroman av Pierre Loti.

Det är min ökenromantik.[1]

Lindqvist tog med idén hem och år 2002 arrangerade Svenska Akademien ett seminarium om hur man skulle kunna åstadkomma ett bibliotek med svenska klassiker, fritt tillgängligt för alla på internet. Det ledde till att Johan Svedjedal fick uppdraget att utreda förutsättningarna för en webbplats för de centrala delarna av svensk litteratur. I sin rapport drog han upp linjerna för en sådan portal, och han föreslog namnet Litteraturbanken eftersom den skulle knytas till Språkbanken vid Göteborgs universitet, en väletablerad institution för språkteknologi och korpuslingvistik. Litteraturbanken skulle fokusera på skönlitteratur och humaniora, samt vara mer utåtriktad och lättillgänglig för en bred grupp användare.

Riksbankens Jubileumsfond finansierade ett pilotprojekt åren 2004–2005, därefter blev Litteraturbanken en ideell förening. Verksamheten har huvudsakligen bekostats av Svenska Akademien, medan Vitterhetsakademien flera gånger gått in med riktat stöd för särskilda ändamål. Också Svenska Litteratursällskapet i Finland har givit finansiellt stöd, de övriga medlemmarna har stöttat verksamheten med material och resurser.

Vilken funktion skulle då en svensk litteraturbank ha? Det fanns ju andra webbplatser för svensk litteratur, inte minst Projekt Runeberg (runeberg.org) som har digitaliserat litteratur sedan 1992. Dramawebben (dramawebben.se) har tillkommit på senare år. Google Books har också skannat in en stor mängd svensk skönlitteratur, helt enkelt genom de svenska samlingarna i de amerikanska och tyska bibliotek som hittills avverkats. Det är mycket värdefulla resurser, och Litteraturbanken är inte avsedd att konkurrera med dem utan har en annorlunda inriktning.

Litteraturbankens strävan har inte i första hand varit att bli ett mer eller mindre allomfattande arkiv, utan att bli en plats där lärare, forskare, studenter och alla andra kan uppehålla sig för arbete, studier eller avkoppling.

Litteraturbanken vill

- erbjuda en uppsättning material som omfattar inte bara källtexter utan också moderna vetenskapliga utgåvor, nyskrivna presentationer och inledningar till litterära sammanhang;
- erbjuda materialet i noga kontrollerad form;
- bli en plats för arbete, studier och förströelse;
- utveckla tekniska lösningar för vetenskaplig utgivning och erbjuda stöd för sådana projekt. I första hand har Litteraturbanken samarbetat med Selma Lagerlöf-arkivet, som nu införlivats i webbplatsen.

Sådana ambitioner tar tid att förverkliga, och Litteraturbanken har dessutom på senare år utvidgat ambitionerna till att

- stödja lärare i skola och gymnasium med material, didaktiska introduktioner, lektionsidéer och uppgifter så att lärare kan sända sina elever dit. Våren 2014 lanserades därför Litteraturbankens skola, som kommer att vidareutvecklas de närmaste åren.
- utveckla storskaliga forskningsmetoder som öppnar upp litteraturen som vetenskapligt källmaterial på allvar. Detta arbete befinner sig på förberedelsestadiet.

Litteraturbanken riktar sig alltså mot ett antal olika målgrupper, men det finns skäl för den breda hållningen: ingen annan tar ansvar för att den svenska skönlitteraturen blir fritt och brett tillgänglig för allmänheten i det nya digitala samhället. Det är viktigt att det materialet förs in i skolorna på ett sätt som ger de unga chansen att bli bekanta och förtrogna med äldre texter, trots att dessa först verkar mycket avlägsna. Det är viktigt att folkbiblioteken får tillgång till fria e-böcker som de kan

erbjuda sina användare. Det är viktigt att forskare och lärare på universiteten får tillgång till pålitliga versioner av källtexter och vetenskapliga utgåvor, både för att förenkla arbetet och för att möjliggöra nya forskningsuppgifter. Allt detta, tror vi, skapar förutsättningar för modern forskning på äldre material och på vårt kulturarv.

Vid utgången av 2014 erbjöd litteraturbanken.se cirka 1 200 verk eller 275 000 boksidor av 600 författare. Verken är tillgängliga för läsning på webbplatsen, de flesta också för sökning. Drygt 600 av böckerna fanns också som epub-filer för nerladdning till mobiltelefoner och läsplattor eller datorer. Det mesta materialet är sådant som inte längre är rättighetsskyddat, det vill säga att det har gått mer än 70 år sedan författarens död, men här finns också verk av Katarina Frostenson, Sven Lindqvist, Stig Larsson och andra samtida författare. Urvalet är i hög grad svenska klassiker och andra viktiga verk som kan vidga vår förståelse av kulturarvet. Litteraturbanken erbjuder också introduktioner och presentationer som kan leda användare in i litteraturskatten – barnbokens historia i Sverige, deckarens, skräckberättelsens utveckling och så vidare, med utvalda illustrativa verk. Där finns en mängd vetenskapliga utgåvor och inledningar av mer akademiskt slag, samt Litteraturbankens skola med lektioner för lärare och elever i skolan och gymnasiet (litteraturbanken.se/skola).

Allt är fritt tillgängligt för den nyfikna läsaren som vill hitta gamla vänner eller söka sig fram till nya bekantskaper. Det är öppet för lärare att styra elever och studenter till verk som inte är så lättillgängliga på andra sätt. För forskarna är avsikten att erbjuda ett rikt och pålitligt material, både källor, vetenskapliga utgåvor och facklitteratur – allt detta är under uppbyggnad.

Litteraturbanken uppdateras månatligen med nya böcker och sänder i samband med det ut sitt nyhetsbrev. Antalet besök är drygt 1 000 om dagen, men hur länge gästerna stannar varierar kraftigt. Statistiken blir också mindre och mindre rättvisande, för en allt viktigare funktion har

blivit att erbjuda epub-filer som användaren hämtar hem en gång och sedan kan sprida till andra. Denna funktion har blivit viktigare allt eftersom den digitala bokmarknaden utvecklats och bibliotekens behov av digitala böcker för utlåning har ökat.

Forskningsfrågor

Litteraturbanken tillhandahåller pålitliga versioner av ett stort forskningsmaterial, inte minst vetenskapliga utgåvor av Selma Lagerlöf, August Strindberg (de textkritiska kommentarerna publiceras endast i Litteraturbanken) och C.J.L. Almqvist, och samtliga utgåvor av Svenska Vitterhetssamfundet och Svenska Fornskriftsällskapet (under produktion). Allt detta kommer till användning för traditionella forskningsuppgifter och man har dessutom möjlighet att göra sökningar i större eller mindre urval av materialet.

Men de verkligt intressanta forskningsfrågorna är de som hittills har kunnat bedrivas internationellt, där man redan har byggt upp väldiga mängder digitaliserad litteratur. I synnerhet den engelskspråkiga forskningen har tillgång till stora korpusar med många tusen böcker, och internationellt har man kunnat se en tydlig förändring i inriktning de senaste åren. Tidigare var man inriktad på att digitalisera så mycket som möjligt, men nu har man så stora materialmängder att man i allt högre grad börjat fokusera på hur de materialen effektivast kan utnyttjas för forskningsändamål. Fältet *digital humaniora* är stort, men inte minst har man utvecklat en rad spännande tekniker för så kallad *text mining* – man ser databaserna som gruvor där man kan bedriva industriell utvinning. Det kan låta avskräckande, men det öppnar upp helt nya möjligheter. Metoderna är kvantitativa, och har väckt många frågor om huruvida humaniora är på väg mot en restriktiv positivism. Men de kvantitativa metoderna är sällan nog i sig själva, fördelen med dem är i stället att man kan använda dem som språngbräda och korrelat för kvalitativa undersökningar. Just föreningen mellan nya kvantitativa och kvalitativa metoder är en stor utmaning, och en stor potential för framtiden.

Det är inte så märkligt att metoderna ger ett industriellt intryck, men det beror till viss del också på en förskjutning av den tekniska bearbetningen som inte alls är negativ. Så länge man hade mindre digitaliserade material kunde man vårda dem med kärlek och omsorg: märka upp dem på olika sätt. Inom lingvistiken har man länge märkt upp texter med ordklasser, satsdelar med mera, och på andra områden har man kunnat tillföra information om teman, vem som talar, och så vidare. Men med så stora textmängder som det nu handlar om, har det centrala varit att utveckla tekniker att utforska råa, ”smutsiga” material. När det gäller tusentals böcker handlar det alltså om att utveckla metoder att bearbeta dem utan mer än absolut nödvändig handpåläggning, och att i stället göra programmen bättre på att leta fram det man är ute efter.

Tematisk utvinning och analys

Fördelarna med det angreppssättet är avsevärda. Den *Franska encyklopedin* är ett aktuellt exempel. Den utgavs i 28 band av Denis Diderot och Jean d’Alembert åren 1751–1765 och var redan från början ett problem därför att den med sina drygt 70 000 artiklar rymde enorma mängder kunskap men bara få vägar att nå fram till den kunskapen. Den var alfabetiskt ordnad på uppslagsord, men hur skulle man få överblick över alla artiklar som hörde samman, som behandlade likartade ämnen? Redaktörerna Diderot och d’Alembert hanterade problemet efter bästa förmåga, genom att upprätta index som gjorde det möjligt att söka sig fram till det man var ute efter. Man kunde alltså få index över artiklar som pekade mot kategorier som naturrätt (*Droit naturel*), geografi (*Géographie*), grammatik av ett slag som rymde en mängd ordboksdefinitioner (*Grammaire*), moral (*Morale*) och teologi (*Théologie*). Det var till stor hjälp, men överblicken blev fortfarande provisorisk. Kategorierna blev mycket stora, men ändå långt ifrån kompletta. Digitala metoder ger helt nya ingångar, som professorn i franska vid University of Chicago Robert Morrissey och hans forskargrupp har visat. De använder en metod för utvinning och utformning av teman, kallad *topic modeling*: ett sätt att

med maskinens hjälp urskilja tematiska samband i stora textmaterial. Man låter maskinen själv avgöra vad som är ett ”tema” genom att analysera hur ord förekommer i samma sammanhang. Programmet organiserar samförekomsterna av ord i grupper och kartlägger hur sådana teman fördelar sig över hela det stora materialet.

När Morrissey lät programmet analysera hela *Franska encyklopedin* fick maskinen fram teman som i hög grad motsvarade rubrikerna Diderot själv hade valt – naturligt nog, en mängd artiklar var ju skrivna inom dessa teman. Men vad maskinen kunde visa var att de olika ämnena behandlas på långt fler ställen än vad som framgår av Diderots register. Ämnet *naturrätt*, till exempel, kan givetvis behandlas utan att orden *droit* eller *naturel* skrivs ut. Programmet identifierade en rad ord som förekommer i diskussioner kring naturrätt, som *droit loi nature société hommes raison choses état justice naturel juste vie gens devoirs morale vertu souverain*, och pekade ut en lång rad passager där naturrätt diskuteras eller frågor som angår den berörs, men under helt andra kategorier än Diderots index visar. Det gällde 61 artiklar som sorterats till *Grammaire* och 25 artiklar som sorterats till *Géographie*, till exempel.

Programmet fångade alltså upp dolda bearbetningar av temat: det är viktigt i sig för att precisera förståelsen av hur naturrätt faktiskt behandlas i den *Franska encyklopedin*, men programmet avslöjade också en intressant subversiv tendens i verket. Det pekade ut artikeln *Inviolable*, ’okränkbar’ (sannolikt skriven av Diderot själv), där man kan läsa något som inte står någon annanstans: ”La liberté de conscience est un privilege inviolable.” Åsiktsfrihet definieras som en okränkbar rättighet – och dessutom här som del av den oifrågasättliga naturrätten, på ett sätt som inte fick plats i någon huvudartikel och som inte var förenligt med den politiska situationen i 1750-talets Frankrike. Passagen hade inte gått att söka fram med konventionella metoder, utan framträder på grund av ord som associeras med varandra på svårfångade sätt – och som programmet fångar upp. På samma sätt spårade programmet upp artikeln *Supplanter* (’ersätta’ eller ’konkurrera ut’), där tyranni och därmed en-

välde beskrivs som en onaturlig styresform, också ett mycket kontroversiellt yttrande som bara kunde smygas in.[2]

Topic modeling, tematisk utvinning och utformning, ger alltså möjligheten att i stora material se samband som inte är manifesta, och som dessutom är fördelade över så stora mängder av texter att ingen enskild forskare kan överblicka dem. Om man jämför den här metoden med tidigare humanistisk och samhällsvetenskaplig forskning, så kan man iaktta en intressant förskjutning av det subjektiva momentet i forskningen. Normalt definierar forskaren frågor och kategorier som sedan anläggs på materialet: då ligger ett subjektivt moment redan i formuleringen av frågor. Med hjälp av metoder som *topic modeling* kan man delvis motverka det subjektiva momentet: de teman som programmet urskiljer är baserade helt på hur ord förekommer tillsammans, i samma kontext. Forskarens analys av materialet kommer sedan att innehålla subjektiva moment, men de förskjuts alltså längre framåt i processen.[3] Och möjligheten uppstår att maskinerna bjuder på helt nya insikter och uppslag till det som inte gick att tänka på förhand. Samtidigt finns det förstås mycket att vara vaksam mot: det är centralt att förstå hur programmen fungerar, vad de förbigår och hur de viktar resultaten.

Preciserad tematisk utvinning och analys

Många skandinaviska böcker har blivit digitaliserade, inte minst av Google Books – de böcker som har funnits i de amerikanska och tyska bibliotek Google Books har upprättat avtal med. Det är svårt att få tillgång till de böckerna för forskning, men två amerikanska skandinavister, Peter Leonard, Librarian for Digital Humanities Research vid Yale och Timothy Tangherlini, professor vid University of California, Los Angeles, fick tillgång till alla danska böcker från det sena 1800-talet som Google hade digitaliserat. De utvecklade en egen version av *topic modeling*, som visade sig mycket fruktbar. Den går ut på att man utvinner teman från ett eller några verk som man är väl bekant med. Man får då ett antal teman som man kan förfina och välja mellan, och därefter

”transplanterar” man sökningen på det större materialet så att man får se hur dessa teman återkommer och fördelar sig över tusentals böcker. Leonard och Tangherlini valde de första danska översättningarna av Darwins *On the Origin of Species* och *The Descent of Man*. De publicerades 1872 och 1875: översättaren var J.P. Jacobsen, själv en tongivande skönlitterär författare. De teman som urskildes var förstås inte överraskande i sig: det kanske tydligaste temat kunde sammanfattas ”social instinkt” och inbegrep ord för samhälle, handling, känsla, moral, dygd. Ett annat framträdande tema rörde ”kampen för överlevnad”.

Det intressanta uppstod när Leonard och Tangherlini överförde resultaten på det stora materialet. Först fick de ett stort antal träffar på välkända manliga författare, de så kallade *Moderna genombrottets män* som vi vet förenade naturalism med samhällsengagemang och – darwinska teorier. Men de fick också många träffar i andra författarskap, idag bortglömda, många av dem kvinnor. Det här gav underlag för en betydligt mer nyanserad bild av litteraturhistorien. Ur ett bredare samhälleligt perspektiv intressantare var fynden de gjorde i tidskrifter och facklitteratur, där debattörer under inflytande från Darwin började förfäkta tanken att kriminalitet är ett symptom på sociala sjukdomar och måste behandlas, snarare än bestraffas. Datorprogrammet visar oss här mot början till en lång rad fängelsereformer och samhällsdiskussioner. Inte minst givande visade sig kopplingarna till historieskrivningen vara: programmet klargjorde hur man plötsligt började beskriva händelser ur Danmarks historia i darwinska termer.[4] Ett steg, kan man säga, mot att följa hur man började skriva om Danmarks historia – ett allra första steg, skall understrykas.

Det finns många falluckor i detta slags material och metoder. Materialet blir aldrig fullständigt, och det är svårt att bedöma hur representativa de resultat man får faktiskt är. Ett påtagligt problem är kvaliteten på textigenkänningen: ju äldre böcker, desto fler fel uppstår och de kan bli så stora att programmet faktiskt inte känner igen orden. Resultaten har alltså klara begränsningar och det rent kvantitativa får inte komma

att dominera frågeställningar och tolkning. Men de digitala materialen och metoderna ger remarkabla möjligheter att mer effektivt bearbeta gamla problem, och att ställa upp helt nya forskningsfrågor. Den här typen av forskning är alls inte begränsad till skönlitteratur. Men den har potentialen att göra skönlitteraturen till ett fruktbart fält för en lång rad humanistiska och samhällsvetenskapliga studier: de tvärvetenskapliga samarbetena är väl etablerade inom det vida fält som brukar kallas digital humaniora. I mötet kring digitala material och metoder uppstår nya möjligheter för de olika disciplinerna att utbyta perspektiv, erfarenheter och metoder, och utveckla nya forskningsfrågor. Inte minst kombinationen mellan kvantitativa och kvalitativa metoder är en utmanande potential.

Vi får möjligheten att följa utvecklingen på många områden och sätta in den i kronologiska och kausala sammanhang. Antalet tänkbara frågeställningar är stort: de kan gälla vilka samhällsfrågor som problematiseras, hur frågor kring världsbild, kön, subjektet, nationell identitet eller det främmande bearbetas, vilka konsumtionsvanor som kommer till uttryck, hur ord och begrepp förändrar sina innebörder eller vilka estetiska föreställningar som utprovas i teori och praktik. Exemplen kan mångfaldigas: i skönlitteraturen pågår ständiga förhandlingar mellan gamla och nya värderingar. Ett väsentligt perspektiv som börjat anläggas av historiker är prosafiktionen inte bara som spegling av samhället utan också som påverkan på samhället: det kan vara konsumtionsmönster, normer, föreställningar av olika slag som inte bara tas upp i litteraturen utan också sätter sin prägel på mottagarna.[5] Med hjälp av digitala resurser blir det möjligt att systematiskt kartlägga sådana företeelser i skönlitteraturen: kan man sedan också ställa skönlitteraturen mot tidskrifter, tidningar, facklitteratur, utredningar, protokoll blir utväxlingen än större: digitala material och metoder har stora potentialer på en lång rad områden.

Digitaliseringsproblemet

I Sverige finns utomordentlig språkteknologisk expertis för detta slags tekniker, och den kommer att stärkas ytterligare genom det svenska samarbetet inom CLARIN (Common Language Resources and Technology Infrastructure: http://sweclarin.se). Men för svenska forskare finns ett mycket påtagligt problem. Svenska politiker talar gärna om att Sverige skall vara världsledande i den digitala utvecklingen, men i praktiken ligger Sverige i internationell strykklass när det kommer till digitalisering av böcker. Inom många språkområden har man digitaliserat väldiga material, också i vår direkta närhet. Om ett år kommer varje bok som tryckts på norska att finnas digitaliserad på Nasjonalbibliotekets webbplats bokhylla.no. I Sverige har Kungl. biblioteket förväntats hantera digitaliseringen inom befintlig budget, en omöjlig uppgift förstås. Från statligt håll har medel för digitalisering främst ingått i arbetsmarknadsåtgärder som inte har resulterat i sammanhållna, konsekventa databaser. Vi har alltså inga stora litterära material att bedriva den här sortens forskning på. Litteraturbanken börjar komma upp i en sådan volym att man kan göra vissa undersökningar, och vi har etablerat ett systerprojekt där prosafiktion från 1800-talet etablerats efter kriterier som baseras på kriterier som siktar till en högre grad av representativitet (spf1800-1900.se) men materialet är ändå jämförelsevis litet. Det behövs storskaligt och konsekvent genomförd digitalisering för att vi fullt skall kunna utnyttja de nya möjligheterna.

Situationen har krävt stor tålmodighet, ända fram till nu. I en ny arbetsmarknadsåtgärd har Svenska Migrationscentret fått i uppdrag att inrätta ett stort antal anpassade tjänster för arbetslösa för digitalisering av material som angår migration och integration, i första hand arkiv kring till exempel Amerikaemigranterna. Migrationscentret och jag har enats om att skönlitteraturen, och det svenska bokmaterialet i stort, är ett omistligt material för förståelsen av hur människor har förflyttat sig, uppfattat sig själva och det främmande, införlivat eller reagerat mot an-

dra kulturer. Jag har beskrivit ett projekt där 20 personer i fem år arbetar efter Litteraturbankens direktiv med att digitalisera all skönlitteratur av svenska författare från 1700 till 1957 som finns i Göteborgs universitetsbibliotek, och tillgängliggöra allt som inte skyddas av rättigheter. Det rör sig om 20–25 000 böcker.

Om allt går enligt planerna kommer Litteraturbanken år 2020 – vid sidan av sina noggrant kontrollerade böcker, introduktioner och presentationer – att kunna erbjuda ett mycket omfattande, konsekvent strukturerat material. Ambitionen är att också skapa ett nationellt samarbete där fler forskningsbibliotek utvecklar motsvarande projekt men med andra urval – tidskrifter, 1600-talets tryck med mera. Med sådana material finns mycket goda förutsättningar för en högst produktiv tvärvetenskaplig forskning och metodutveckling. I anslutning till Alvin, den plattform för olika system för kulturarvsmaterial som utvecklas av en rad universitetsbibliotek, kommer många nya möjligheter att öppnas.

Föredrag den 7 oktober 2014

NOTER

1. Sven Lindqvist, *Ökendykarna*, 1990, 135–136.
2. Glenn Roe, Clovis Gladstone & Robert Morrissey, ”Discourses and Disciplines in the Enlightenment: Topic Modeling the French *Encyclopédie*”, *Digital Humanities 2014. Book of Abstracts*, 336–338.
3. Se vidare John W. Mohr & Petko Bogdanov, ”Introduction—Topic models: What they are and why they matter”, *Poetics* 41, 2013, 545–569.
4. Timothy R. Tangherlini & Peter Leonard, ”Trawling in the Sea of the Great Unread: Sub-corpus topic modeling and Humanities research”, *Poetics* 41, 2013, 725–749.
5. Sådana initiativ finns till exempel i *Historier. Arton- och nittonhundratalens skönlitteratur som historisk källa*, utg. Christer Ahlberger et al., Göteborg 2009, och *Moderna historier. Skönlitteratur i det moderna samhällets framväxt*, utg. Henric Bagerius & Ulrika Lagerlöf Nilsson, Lund 2011.